



King's Research Portal

DOI:

[10.1109/LCOMM.2016.2518684](https://doi.org/10.1109/LCOMM.2016.2518684)

Document Version

Peer reviewed version

[Link to publication record in King's Research Portal](#)

Citation for published version (APA):

Ghoreishi, S. E., & Aghvami, A-H. (2016). Power-efficient QoE-aware Video Adaptation and Resource Allocation for Delay-constrained Streaming over Downlink OFDMA. *IEEE COMMUNICATIONS LETTERS*, 20(3), 574 - 577. <https://doi.org/10.1109/LCOMM.2016.2518684>

Citing this paper

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Power-Efficient QoE-Aware Video Adaptation and Resource Allocation for Delay-Constrained Streaming over Downlink OFDMA

Seyed Ehsan Ghoreishi, *Student Member, IEEE* and A. Hamid Aghvami, *Fellow, IEEE*

Abstract—In this letter we propose a QoE-aware video adaptation and resource allocation approach for power-efficient streaming over downlink OFDMA systems. Our adaptation scheme selectively drops packets from a video stream to produce a lower bit-rate version under QoE and delay constraints. This results in a reduction of the load and an increase of the video capacity of the wireless network. Our resource allocation target is to minimize the transmit power by considering the delay requirements of each stream identified in the video adaptation phase. Experimental results have shown significant performance enhancement of the proposed system in terms of end-to-end delay and power efficiency while satisfying QoE requirements.

Index Terms—Video adaptation, QoS-QoE mapping, scalable video coding (SVC), resource allocation.

I. INTRODUCTION

THE extensive growth in the adoption of smartphones and tablets has led to a continuous increase in mobile video traffic. This new phenomenon has urged mobile operators to redesign their networks to support more simultaneous video streams while maintaining stringent delay bounds and guaranteeing a certain level of QoE for individual users.

By providing multiple source video bit-rates for a single video, adaptive bitrate streaming (ABR) increases the wireless network capacity to serve more video requests concurrently [1]. Furthermore, many studies have proposed wireless video caching as a way to maximize the video capacity of wireless networks while enhancing the user-perceived QoE [1]–[4].

With ABR streaming, each video is divided into multiple chunks and each chunk can be requested at different bit-rates. Therefore, for an entire video to be served from the cache, one can cache all rate variants. However, this significantly increases backhaul and storage requirements as a video could be encoded into more than 40 versions to meet the heterogeneity of user devices and network conditions [5]. Moreover, the available transmission rate in wireless channels is time-varying and hard to predict. Hence, the selected bitstream transmitted by a content server distant from the user may not match the user's transmission characteristics [3].

Alternatively, we can cache only the best quality video and use a processing resource to perform transrating [1], [3]. However, it consumes tremendous computing and storage resources to encode videos into different bit-rates in real-time and store the encoded streams [4].

The authors are with the Centre for Telecommunications Research, King's College London, London WC2R 2LS, U.K. (e-mail: {seyed_ghoreishi, hamid.aghvami}@kcl.ac.uk).

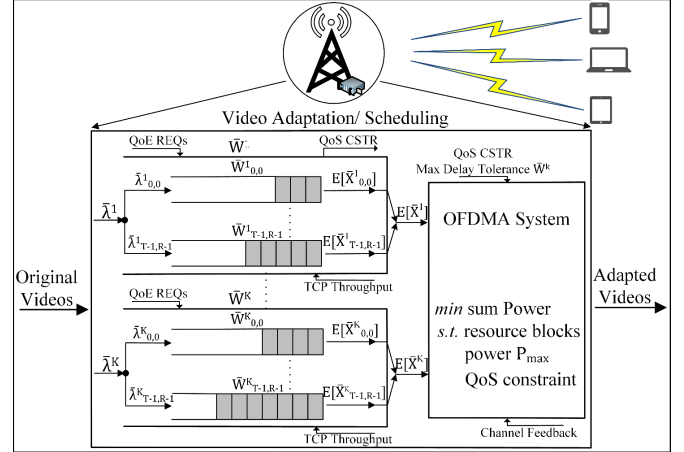


Fig. 1: Video adaptation/ scheduling system at network edge.

To decrease radio access network (RAN) backhaul and cache storage requirements of downloading and caching multiple bit-rate versions of a video, as shown in Fig. 1, we enhance our RAN with a queuing-based video adaptation/ resource allocation (RA) module. A delay-constrained scalable video coding (SVC)-specific active queue management technique which transrates a stream to a lower bit-rate and leads to a power-efficient RA has not been investigated and hence is the focus of this letter. It drops packets that have minimal negative impact on the user's QoE to satisfy a certain level of QoE for a user. This in turn reduces network load and delay, and increases the capacity to serve more concurrent streams.

II. SYSTEM MODEL

We focus on the downlink of LTE networks and consider a single-cell multi-user scenario as shown in Fig. 1. The system consists of K mobile users (video streams) indexed by the set $\mathcal{K} \triangleq \{1, \dots, k, \dots, K\}$, sharing L resource blocks (RBs) indexed by $\mathcal{L} \triangleq \{1, \dots, l, \dots, L\}$ in an OFDMA cell. The channel is assumed to be frequency-selective Rayleigh fading, with flat fading within each RB. Each H.264/SVC stream has a number of temporal layers and quality layers. We index temporal and quality layers of stream k by $\mathcal{T} \triangleq \{0, \dots, t, \dots, T^k - 1\}$ and $\mathcal{R} \triangleq \{0, \dots, r, \dots, R^k - 1\}$, respectively.

We deploy a statistical queuing model to express the delay limitation of a stream with an equivalent cross-layer constraint. Therefore, as in [6], we assume that packets arrive to each user k 's buffer q^k based on a Poisson arrival process. Within

q^k , the system places packets from the r^{th} quality layer of temporal layer t of sequence k into virtual queue (VQ) $q_{t,r}^k$, which follows the dynamics of M/G/1 queues [7]. The arrival rates in M/G/1 are Poisson processes, which are highly suitable for modeling SVC video traffic [8]. Moreover, the service time can follow any general statistical distribution. This is due to the fading channel, which makes the service process hard to model [6]. As shown in Fig. 1, we describe the parameters of q^k and $q_{t,r}^k$ by characteristic tuples $[\bar{\lambda}^k, \mathbb{E}[X^k], \bar{W}^k]$ and $[\bar{\lambda}_{t,r}^k, \mathbb{E}[X_{t,r}^k], \bar{W}_{t,r}^k]$, respectively. $\bar{\lambda}_{t,r}^k, \mathbb{E}[X_{t,r}^k]$ and $\bar{W}_{t,r}^k$ are the arrival rate, service time and waiting time of the packets at $q_{t,r}^k$, and $\bar{\lambda}^k, \mathbb{E}[X^k]$ and \bar{W}^k are those of the packets at q^k .

Dropping packets from the VQs decreases the queuing delay and congestion in the network. It also increases the network capacity (the number of concurrent video requests that can be served). However, packet loss causes a certain reduction in the user QoE of a video depending on the importance of the video layer containing the dropped packet. This is estimated using the QoE metric proposed in Section II-A. Thus, we drop packets from different layers of a stream and produce a lower bit-rate stream that satisfies the user's QoE requirements.

We formulate the video adaptation problem as minimization of the queuing delay of user streams by means of dropping packets under QoE provisioning. The power-efficient RA OFDMA module then uses the calculated optimal queuing delay (which takes the QoE requirements and decoding deadline of the videos into account) as a constraint that specifies the maximum delay tolerance for the videos. The OFDMA module transforms this delay constraint into a cross-layer constraint for OFDMA systems using the method proposed in Section II-C and finds the optimal RB and transmit power allocation policies \mathbf{P}^* and \mathbf{x}^* , respectively to satisfy this constraint.

In the following sections, we explain the proposed QoE metric model. It provides a relationship between packet loss ratio and reduction in QoE. We then formulate a relationship between the loss ratio at a queue and queuing delay. This leads to a relationship between user QoE and packet loss ratio. Next, we transform the queuing delay requirements into a cross-layer constraint by formulating a relationship between the average data rate of a stream and its delay threshold. Lastly, we formulate the video adaptation and RA problems.

A. QoE Metric Model

We use the multi-scale structural similarity (MS-SSIM) index [9], which provides a good approximation of user-perceived quality. It calculates relative quality scores between a reference video frame and a distorted version. We use the QoS-QoE mapping technique proposed in [10]. It interprets packet loss ratio into a system-level QoE measure. We calculate the degradation in QoE caused by data drops at each video layer. Thus, for a given video stream, we perform a Monte Carlo simulation where we drop a fixed percentage ($\rho_{t,r}^k$) of packets from each temporal/quality layer uniformly at random. We estimate the average QoE achievable $\mathbb{E}[q(\rho_{t,r}^k)]$ when the packet loss ratio in a temporal/quality layer is $\rho_{t,r}^k$. At each run, the video is decoded and the quality index is measured. For each $\rho_{t,r}^k$ value, we perform different instances of the test

to find the average quality for $0 \leq \rho_{t,r}^k \leq 1$. Repeating over all temporal/quality layers, we obtain the empirical mapping.

Proposition 1. QoE reduction at stream k is defined as [10]

$$\mathcal{D}_{\text{overall}}^k = (1 - \alpha^k) q_{\text{max}}^k = \sum_{t=0}^{T^k-1} \sum_{r=0}^{R^k-1} \mathcal{D}(\rho_{t,r}^k), \quad (1)$$

where q_{max}^k is the quality in the absence of losses for stream k , α^k is the fractional quality degradation due to packet loss, and $\mathcal{D}(\rho_{t,r}^k) = q_{\text{max}}^k - \mathbb{E}[q(\rho_{t,r}^k)]$ is the QoE degradation caused by packet loss ratio $\rho_{t,r}^k$ in temporal layer t , quality layer r .

Proof. The proof directly extends from [10]. \square

In case of TCP-based ABR, we quantify the loss visibility of packets from each video layer over time using the ACK history, as in [10]. After a group of pictures (GoP) is transmitted and its complete ACK history is fed back to the transmitter, a replica of the decoded GoP is reconstructed with the losses from each layer. Then, the corresponding packet loss is computed directly from the ACK history to estimate the channel distortion effects on each video layer.

B. MAC-Layer Modeling from a Cross-Layer Perspective

The average length of M/G/1 queue $q_{t,r}^k$ is given by [7]

$$\bar{L} = \bar{\lambda}^2 \mathbb{E}[X^2] / 2(1 - \bar{\lambda} \mathbb{E}[X]), \quad (2)$$

where $\bar{L}, \mathbb{E}[X], \mathbb{E}[X^2]$ and $\bar{\lambda}$ are used to denote $\bar{L}_{t,r}^k, \mathbb{E}[X_{t,r}^k], \mathbb{E}[X_{t,r}^k{}^2]$ and $\bar{\lambda}_{t,r}^k$, respectively. $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ are the first and second moments of the service time at queue $q_{t,r}^k$.

We can estimate the average arrival rate of $q_{t,r}^k$ by $\bar{\lambda}_{t,r}^k = \bar{s}_{t,r}^k (n_{t,r}^k / N^k) f^k$ [11], where $\bar{s}_{t,r}^k$ is the average size of a video frame in temporal layer t of the r^{th} quality layer, N^k is the number of frames in a GoP and f^k is the frame rate of stream k . $n_{t,r}^k$ is the number of frames in the t^{th} temporal layer of each quality layer, which can be derived from [12]

$$n_{t,r}^k = \begin{cases} 1 & \text{if } t \in \{0, 1\} \\ 2^{t-1} & \text{if } 2 \leq t \leq \log_2 N^k. \end{cases} \quad (3)$$

Based on the Little theorem [7], the average waiting time in each queue is $\bar{W}_{t,r}^k = \bar{L}_{t,r}^k / \bar{\lambda}_{t,r}^k$, where $\bar{L}_{t,r}^k = (1 - \rho_{t,r}^k) \bar{L}_{t,r}^k$ is the average queue length in the presence of packet loss ratio $\rho_{t,r}^k$ in $q_{t,r}^k$. Therefore, substituting (2) and $\bar{L}_{t,r}^k$ into $\bar{W}_{t,r}^k = \bar{L}_{t,r}^k / \bar{\lambda}_{t,r}^k$, the average waiting time in an M/G/1 queue is

$$\bar{W} = (1 - \rho) \bar{\lambda} \mathbb{E}[X^2] / 2(1 - \bar{\lambda} \mathbb{E}[X]), \quad (4)$$

where \bar{W} and ρ are used to denote $\bar{W}_{t,r}^k$ and $\rho_{t,r}^k$, respectively.

C. Delay Requirements to Data Rate Transformation

We estimate the maximum delay tolerance \bar{W}_{max}^k in the next section, which puts an upper-bound on the delay experienced by stream k . However, in order to transform this QoS constraint into a cross-layer constraint, using an M/G/1 queuing model, *Proposition 2* formulates a relationship between the average scheduled effective data rate of each user k and \bar{W}_{max}^k .

Proposition 2. A necessary condition to meet a maximum delay of \bar{W}_{\max}^k for a stream k in an OFDMA system is [6]

$$\mathbb{E} \left[\sum_{l=1}^L R_{k,l} \cdot x_l^k \right] \geq \left(\sqrt{\bar{\lambda}^k \bar{W}_{\max}^k (\bar{\lambda}^k \bar{W}_{\max}^k - 2\rho^k + 2)} + \bar{\lambda}^k \bar{W}_{\max}^k \right) \frac{S}{2 \cdot B \cdot t_s \cdot \bar{W}_{\max}^k}, \forall k \in \mathcal{K}, \quad (5)$$

where B is the bandwidth of each RB, t_s is the scheduling slot duration and S is the size of each packet. $\bar{\lambda}^k$ and ρ^k are the average arrival rate and packet loss ratio at q^k . $R_{k,l} = B \log_2(1 + \frac{P_l^k |h_l^k|^2}{\sigma^2})$ is the upper bound on the achievable service rate for user k over RB l , where h_l^k is the channel fading coefficient and σ^2 denotes the noise power.

Proof. The proof extends from [6]. \square

III. VIDEO ADAPTATION AND RESOURCE ALLOCATION

In this section we first formulate the video adaptation as a QoE-constrained queuing delay minimization problem. We derive the optimal packet loss ratio and service rate which minimize the queuing delay and adapt the stream based on the QoE constraint. We then formulate the cross-layer RA problem as a power minimization problem under the delay constraint derived in the video adaptation phase.

A. Optimization Based Video Adaptation/ Scheduling

The objective is to maximize capacity, which we define as the number of concurrent streams that can be served while meeting each stream's QoE and delay requirements. We achieve this aim by minimizing the average queuing delay of each stream, and hence decreasing the queue length in the buffer. This in turn provides a lower bit-rate version of the stream by dropping packets, subject to minimum QoE and maximum decoding deadline constraints at all VQs.

$$\text{minimize}_{\rho, \mathbb{E}[X]} \mathbb{E} \left[\sum_{t=0}^{T^k-1} \sum_{r=0}^{R^k-1} \bar{W}_{t,r}^k \right] \quad (6)$$

$$\text{subject to } \bar{W}_{t,r}^k \leq \bar{W}_{\max,t,r}^k \quad \forall k \in \mathcal{K}, \forall t \in \mathcal{T}, \forall r \in \mathcal{R} \quad (6a)$$

$$\mathcal{D}_{\text{overall}}^k \leq \mathcal{D}_{\max}^k \quad \forall k \in \mathcal{K} \quad (6b)$$

$$\sum_{t=0}^{T^k-1} \sum_{r=0}^{R^k-1} \mathbb{E}[X_{t,r}^k] \leq C^k \quad \forall k \in \mathcal{K}. \quad (6c)$$

The objective function (6) minimizes the average queuing delay of video streams. Constraint (6a) ensures that the average waiting time in each VQ does not exceed the respective average expiry time (decoding deadline) $\bar{W}_{\max,t,r}^k$. The average expiry time of packets in $q_{t,r}^k$ can be adequately approximated by $\bar{W}_{\max,t,r}^k \approx \frac{1}{f^k}$ [13]. Constraint (6b) means that QoE reduction at stream k does not exceed \mathcal{D}_{\max}^k , which is the maximum allowable degradation in the QoE of the stream (decided by operator). In ABR streaming, the requested video rate is adapted to the user's TCP throughput. Therefore, (6c) ensures that sum of the average service rates of stream k 's VQs is upper-bounded by the end-user's TCP throughput C^k .

B. Power-Efficient Delay-constrained Resource Allocation

We deploy the cross-layer RA problem in [6]. It targets to minimize the power transmitted from the base station (BS) to K users while satisfying the delay limitation of each stream derived in Section III-A. The RA problem is formulated as

$$\text{minimize}_{P,x} \mathbb{E} \left[\frac{1}{L} \sum_{k=1}^K \sum_{l=1}^L P_l^k \cdot x_l^k \right] \quad (7)$$

$$\text{subject to } \mathbb{E} \left[\frac{1}{L} \sum_{k=1}^K \sum_{l=1}^L P_l^k \cdot x_l^k \right] \leq P_{\max} \quad \forall k \in \mathcal{K} \quad (7a)$$

$$\sum_{k=1}^K x_l^k \leq 1 \quad \forall l \in \mathcal{L} \quad (7b)$$

$$x_l^k \in \{0, 1\}, P_l^k \geq 0 \quad \forall k \in \mathcal{K}, \forall l \in \mathcal{L} \quad (7c)$$

$$\bar{W}^k \leq \bar{W}_{\max}^k \quad \forall k \in \mathcal{K}. \quad (7d)$$

The objective of the optimization problem in (7) is power and RB allocation in order to minimize the total transmit power in the downlink. P_{\max} in (7a) puts an upper limit on the average total available power at the BS. (7b) and (7c) indicate that each RB can be allocated to one receiver exclusively. We use binary variables x_l^k to represent the RB assignment. (7d) expresses the delay limitation of stream k . \bar{W}_{\max}^k is the maximum delay tolerance for the k^{th} stream, where $\bar{W}_{\max}^k = \mathbb{E}[\sum_{t=0}^{T^k-1} \sum_{r=0}^{R^k-1} \bar{W}_{t,r}^k]$, $\forall k$. $\mathbb{E}[\sum_{t=0}^{T^k-1} \sum_{r=0}^{R^k-1} \bar{W}_{t,r}^k]$ is the optimal solution of problem (6) for stream k . (5) provides a necessary condition for constraint (7d).

A summary of our algorithm is provided in Algorithm 1.

Algorithm 1 Proposed video adaptation/ scheduling algorithm

- 1: Given K streams with properties $T^k, R^k, f^k, N^k, \bar{\lambda}^k$ and maximum allowable QoE degradation \mathcal{D}_{\max}^k ;
 - 2: Use Monte Carlo simulations to estimate $\mathcal{D}(\rho_{t,r}^k)$ for each stream k for $0 \leq \rho_{t,r}^k \leq 1, \forall t \in \mathcal{T}, \forall r \in \mathcal{R}$;
 - 3: Solve (6) to find the optimal packet loss ratio $\rho_{t,r}^{*k}$ of stream k 's temporal/quality layers which produces a lower-rate stream based on \mathcal{D}_{\max}^k ;
 - 4: Obtain the optimal queuing delay $\bar{W}_{t,r}^{*k}$ from (6), which takes the QoE requirements and decoding deadlines of temporal/quality layers of each stream k into account;
 - 5: Calculate the maximum delay tolerance $\bar{W}_{\max}^k = \mathbb{E}[\sum_{t=0}^{T^k-1} \sum_{r=0}^{R^k-1} \bar{W}_{t,r}^{*k}]$ and optimal $\rho_{t,r}^{*k}$ for stream k ;
 - 6: Transform (7d) to a cross-layer constraint using (5);
 - 7: Solve (7) to derive the optimal power P_l^{*k} and RB assignment x_l^{*k} under maximum delay tolerance constraint;
-

IV. NUMERICAL AND SIMULATION RESULTS

In this letter video coding is performed by JSVM 9.19.15. The video sequences "city" (bit-rate ~ 450 kbps) and "foreman" (bit-rate ~ 400 kbps) are used in the simulations. The maximum frame rate is 30 fps and the number of temporal layers and quality enhancement layers are both set to 4. Using the method in Section II-A, we estimate the loss visibility

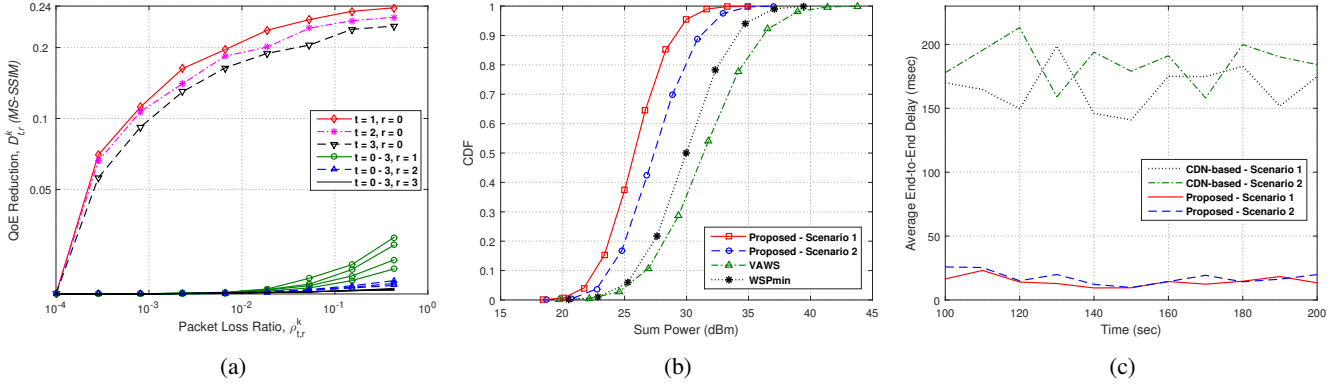


Fig. 2: (a) QoE reduction vs. packet loss ratio for base layers ($t = 1$ to 3 , $r=0$) and enhancement layers ($t=0$ to 3 , $r=1$ to 3) for "city" sequence with y-axis in log scale. (b) CDF of sum power. (c) Comparison of end-to-end delay.

of packets from each video layer. Fig. 2a shows the QoE reduction of "city" sequence (which involves more background motion) when a uniform packet loss is applied to each layer. As shown in Fig. 2a, losses in layers with layer identifier $r = 0$ result in significant degradation in video quality. Due to packet scalability, quality degradation has considerably lower severity when losses occur in upper temporal/quality layers [10].

We now consider the downlink of a single-cell OFDMA system. The bandwidth is 10 MHz (50 usable RBs per TTI). The channel model accounts for Rayleigh fading, large scale path loss and log-normal shadowing. The noise power is -174 dBm/Hz. We assume 8 uniformly distributed users with a minimum distance of 50 m from the eNodeB.

We consider two scenarios in each of which, users have different QoE requirements. In *Scenario 1*, "foreman" video streams are transmitted to the users and each video is adapted dynamically based on the maximum allowable QoE degradation $\mathcal{D}_{\max}^k = 0.3$. In *Scenario 2*, which has higher QoE requirements, we transmit "city" streams and set \mathcal{D}_{\max}^k to 0.1.

We compare our algorithm with WSPmin [14] and VAWS [3] RA schemes. WSPmin minimizes the total transmit power with a minimum rate constraint. In VAWS, RBs are assigned to satisfy minimum rate constraint with the assumption of equal power allocation per RB. It then refines the initial uniform power allocation to ensure that minimum rate requirements are met. It repeats the previous phases to refine power allocation.

The data rate requirements for the users served by WSPmin and VAWS are randomly varying from 100 kbps to 400 kbps as multiples of 50 kbps. Fig. 2b demonstrates the CDF of sum power for different RA schemes generated over 100 iterations using MATLAB. We note that in *Scenario 1*, our proposed scheme outperforms both WSPmin and VAWS algorithms in terms of power efficiency by performing 17.29% better than the former and 24.7% better than the latter in 90% of the times. Likewise, in *Scenario 2*, compared with WSPmin and VAWS, the proposed approach results in 12.37% and 19.81% power-efficiency improvement in 90% of the times, respectively.

Fig. 2c shows a comparison of the proposed approach and the widely used content delivery network (CDN)-based ABR streaming in terms of end-to-end delay using OPNET. Compared with CDN-based streaming where "foreman" videos with \mathcal{D}_{\max}^k set to 0.3 (*Scenario 1*) and "city" videos with

$\mathcal{D}_{\max}^k = 0.1$ (*Scenario 2*) are transmitted to users, the proposed scheme decreases delay by 89.26% and 86.44%, respectively.

V. CONCLUSION

We have proposed a queuing-based video adaptation and RA scheme for video streaming at the network edge. Our approach selectively drops packets from a video stream to produce a lower bit-rate which reduces delay and satisfies a target user QoE. We then allocate resources to meet the delay limitation of the lower rate stream. The results show that our scheme achieves significant performance improvement in terms of reducing delay and power consumption.

REFERENCES

- [1] H. Ahleghagh and S. Dey, "Adaptive bit rate capable video caching and scheduling," in *IEEE WCNC*, Shanghai, China, Apr. 2013.
- [2] N. Golrezaei, A. Molisch, A. Dimakis, and Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, pp. 142–149, Apr. 2013.
- [3] H. Ahleghagh and S. Dey, "Video-aware scheduling and caching in the RAN," *IEEE/ACM Trans. Netw.*, vol. 22, pp. 1444–1462, Oct. 2014.
- [4] G. Gao, W. Zhang, Y. Wen, Z. Wang, and W. Zhu, "Cost-efficient video transcoding in media cloud by leveraging user viewing pattern," *IEEE Trans. Multimedia*, vol. 17, pp. 1286–1296, Aug. 2015.
- [5] Y. Liu, F. Li, L. Guo, B. Shen, and S. Chen, "A server's perspective of internet streaming delivery to mobile devices," in *INFOCOM*, 2012.
- [6] C. C. Zarakovitis, Q. Ni, D. E. Skordoulis, and M. G. Hadjinicolaou, "Power-efficient cross-layer design for ofdma systems with heterogeneous QoS, imperfect CSI, and outage considerations," *IEEE Trans. Veh. Technol.*, vol. 61, pp. 781–798, 2012.
- [7] D. Bertsekas and R. Gallager, *Data Networks*. Prentice Hall, 1992.
- [8] M.-E. Sousa-Vieira, "Suitability of the M/G/∞ process for modeling scalable h. 264 video traffic," in *Analytical and Stochastic Modeling Techniques and Applications*. Springer, 2011, pp. 149–158.
- [9] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *IEEE ACSSC*, Nov. 2003.
- [10] A. A. Khalek, C. Caramanis, and R. Heath, "A cross-layer design for perceptual optimization of H.264/SVC with unequal error protection," *IEEE J. Sel. Areas Commun.*, vol. 30, pp. 1157–1171, 2012.
- [11] H. Wang and G. Liu, "Priority and delay aware packet management framework for real-time video transport over 802.11e WLANs," *Multimedia Tools and Applications*, vol. 69, pp. 621–641, 2014.
- [12] R. Gupta, A. Pulipaka, P. Seeling, L. Karam, and M. Reisslein, "H.264 coarse grain scalable (CGS) and medium grain scalable (MGS) encoded video: A trace based traffic and quality evaluation," *IEEE Trans. Broadcast.*, vol. 58, pp. 428–439, Sept. 2012.
- [13] E. Maani, P. Pahalawatta, R. Berry, T. Pappas, and A. Katsaggelos, "Resource allocation for downlink multiuser video transmission over wireless lossy networks," *IEEE Trans. Image Process.*, vol. 17, 2008.
- [14] K. Seong, M. Mohseni, and J. Cioffi, "Optimal resource allocation for OFDMA downlink systems," in *IEEE Int. Symp. Inf. Theory*, Jul. 2006.